IBM Research

Survey of Sampling Techniques for Data Mining

Sameep Mehta and Vinayaka Pandit

IBM Research – India.

Overview

- First Part: Brief Introduction to Sampling and related issues.

- Second Part: A closer look at Sampling as a technique for data mining taking Association Rule Mining as a case-study.
  - Focus on proving theoretical properties of Sampling Techniques
  - Focus on all relevant calculations so that you see its ease of use and hopefully encourage you to use it in your own work!

First Part

- The slides are developed based on the book– Sampling: Design and Analysis by Sharon L. Lohr
  - Borrowed Examples & Figures from the book

Sampling

- Sample (noun) from Merriam-Webster
  - **:** a representative part or a single item from a larger whole or group especially when presented for inspection or shown as evidence of quality
  - **:** a finite part of a statistical population whose properties are studied to gain information about the whole
  - **:** an excerpt from a musical recording that is used in another artist's recording

- Sampling is useful for
  - Estimation of population characteristics
  - To identify a problem

Some famous example of Sampling

- Shere Hite's survey *Women and Love: A Cultural Revolution in Progress* (1987)
- Widely quoted results
  - 84% of women “not satisfied emotionally with their relationships."
  - 70% of all women “married five or more years are having sex outside of their marriage."
  - 95% of women “report forms of emotional and psychological harassment from men with whom they are in love relationships.“
  - 84% of women report forms of condescension from the men in their love relationships

Criticism

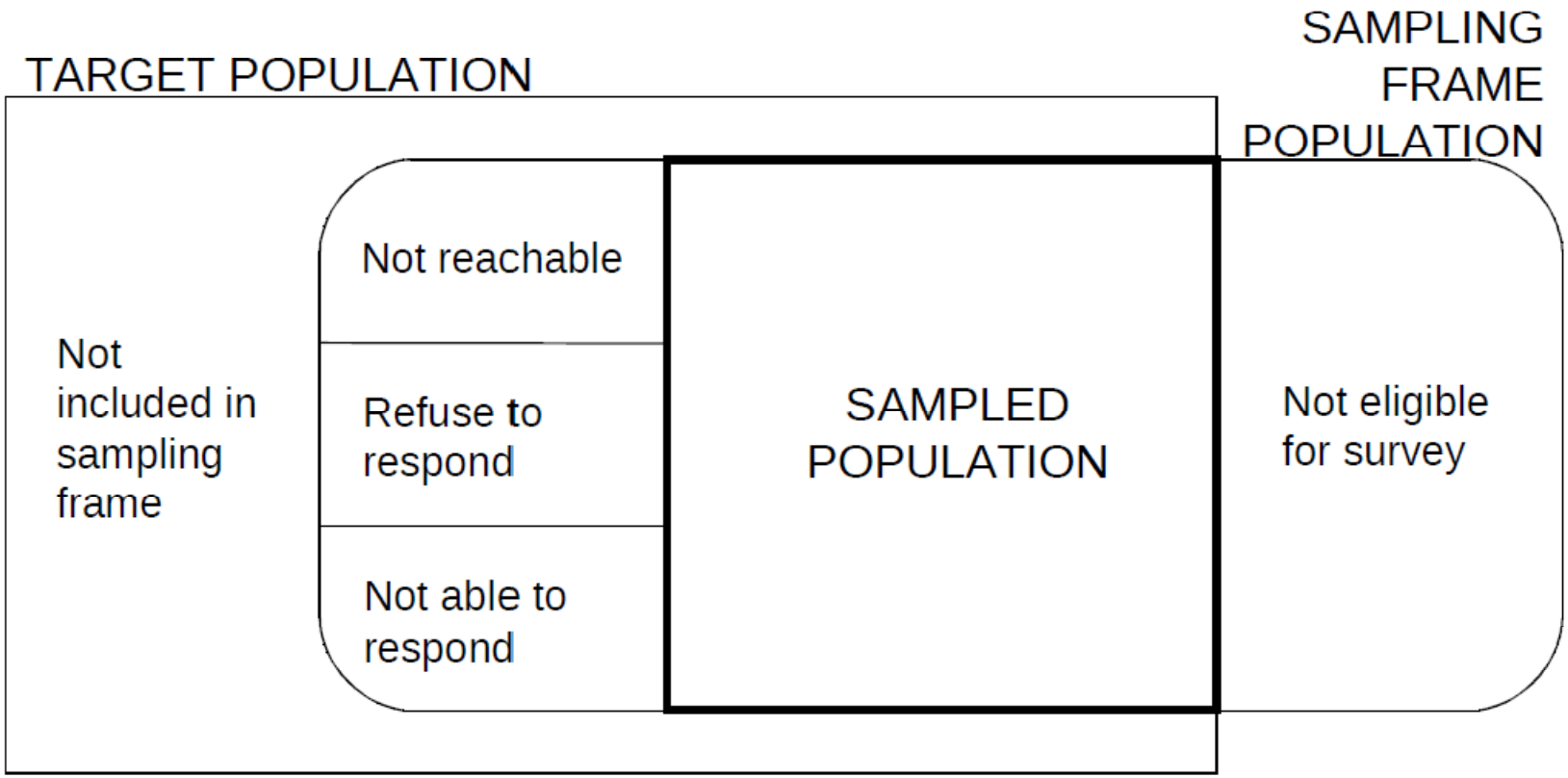
- Sample was self-selected, 4.5% of 100,000 questionnaires were returned
- Questionnaires were mailed to selected organizations –professional women groups, counseling centers, church societies, etc
- Survey had 127 essay questions and many had several parts
- Many questions were vague and misleading
  - Does your husband/lover see you as an equal?
  - Or are there times when he seems to treat you as an inferior?
  - Leave you out of the decisions?
  - Act superior?
- Hites writes "*Does research that is not based on a probability or random sample give one right to generalize from the results of the study to the populatiion at large? If the study is large enough and the sample broad enough and if one generalizes carrfully, yes"* (p 778)
- Most statisticians would answer Hite's question with resounding NO.

Good Samples

- Should be representative of some larger group or population
- Should reproduce the characteristics of interest in the population
  - Example: Movie Magic Town. James Stewart discovered a town Grandview which had exactly same characteristics as United States.
- Some definitions are need to make the notion of a good sample more precise
- Observation unit or element
  - an object on which a measurement is taken
  - Human populations: individuals We collect data on these objects or people
- Target population
  - The complete collection of objects we would like to say something about
  - What we generalize to

Important Concepts

- Sample
  - Subset of the population
  - Sometimes called “subsample" in experimental design
- Sampled population
  - population from which the sample was chosen
  - same as target population?
- Sampling unit
  - Unit actually sampled
  - Can be different from observation unit
  - For example, we may want to study individuals, but do not have a list of all individuals in the target population, so
    - Households as sampling units
    - Individuals as observation units
- Sampling frame
  - List of sampling units


TARGET POPULATION
SAMPLING FRAME POPULATION
Not included in sampling frame
Not reachable
Refuse to respond
Not able to respond
SAMPLED POPULATION
Not eligible for survey

In the Hite (1987) study

- One characteristic of interest was the percentage of women who are harassed in their relationship
  - Element=individual woman
  - Target population=all adult women in the US
  - Sampled population=women belonging to women's organizations who would return the questionnaire
  - Consequently, inferences can only be made to the sampled population, not to the population of all adult women in US

Usefulness of Sample

- Appropriateness to objectives of research
- Accuracy of measurement
- Generalizability of results
- Avoidance of ethical and political problems

Selection Bias

- A part of the target population is not in the sampled population.
- Convenience samples frequently have selection bias.
- Judgment samples frequently have selection bias.
- Misspecifying the target population
- Undercoverage: failure of frame to give access to all elements that belong to target population
- Substitution by convenient member of population
- Non Response
- Allowing the sample of consist solely of volunteers

Example

- Literary Digest took polls and predicted the correct winner in the US presidential race in every election between 1912 and 1932
- They used commercial sampling based on telephone books and auto registration lists
- For the 1936 election they predicted Republican Alf Landon 55% and Democrat Franklin Roosevelt 41%
  - The poll was based on “Ten Million Voters”
- Results: Landon 37%, Roosevelt 61%
- What went wrong
  - Undercoverage: People with telephones and autos were generally more affluent and tended to disapprove of Roosevelt’s economic policies
  - Non response- 10 Million surveys sent. Only 2.3 million received back
- Take home message:
  - Large surveys not necessarily imply correct surveys
  - Design of survey far more important than size of survey

Measurement Bias

- Measurement instrument tends to differ from true value in one direction
- Collection in space
  - Double counting
- People Problem
  - Do not tell truth
  - Forget
  - Underreport bad things
  - Over report good things
  - Questions as not understood
    - 1993 Roper poll reported 25% of Americans did not believe Holocaust happened
    - Reworded question results in 1%
    - Do you own a car?
  - Interviewers Affect
  - Say what interviewers wants to hear

Sampling and Non-Sampling Errors

- Sampling errors – which result from taking a sample instead of examining the whole population
  - Different Sample may give different insights
  - sample-to-sample variation
  - quantified as the margin of error.
- Non-Sampling errors – which cannot be attributed to sample-to-sample variability.
  - Selection bias (e.g. undercoverage, nonresponse)
  - Measurement bias

Why Sampling

- Compared to census:
  - Sampling can provide reliable information at far less cost than a census.
  - Data can be collected more quickly, so estimates can be published in a timely fashion.
  - Unemployment rate for 2005, (2015?)
- Less well known, estimates based on sample survey are often more accurate than those based on a census, because investigators can be more careful when collecting data.
  - Many types of errors can be injected into the census
  - Large administrative organization
  - Many persons in data collection
  - Pressure to produce timely estimates

Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability -- Deming

Simple Probability Samples

- Framework
  - U is the finite population
  - N is the number of units in the population
  - S is a particular sample
  - n is the number of units in S
  - Probability of Selection $\pi_i = \mathrm{P}(\text{unit } i \text{ is included in the sample})$

Sample Distribution

- A fundamental idea in statistics
- Using the data in the sample, we calculate a *statistic*.
  - The distribution of this statistic is the *sampling distribution*.
  - *Random variable* is the set of possible options with associated probabilities.
- Statistic is one realization of a random variable.
- The sampling distribution depends upon
  - the population distribution
  - the sample design.

Example

- Suppose we are interested in the population total.
- Let y_i denote the value of the characteristic of interest for observation unit i.
- $t = \sum_{population} y_i$
- $\hat{t} = N\bar{y}_{sample}$

$N = 4$ example

- $y_1 = 5$; $y_2 = 10$; $y_3 = 15$; $y_4 = 10$
- $t = 5 + 10 + 15 + 10 = 40$
- For $S_1 = \{1, 2\}$, $\hat{t} = 4(7.5) = 30$
- For $S_2 = \{1, 3\}$, $\hat{t} = 4(10) = 40$
- For $S_3 = \{1, 4\}$, $\hat{t} = 4(7.5) = 30$
- For $S_4 - \{2, 3\}$, $\hat{t} - 4(12.5) - 50$
- For $S_5 = \{2, 4\}$, $\hat{t} = 4(10) = 40$
- For $S_6 = \{3, 4\}$, $\hat{t} = 4(12.5) = 50$

- The sampling distribution of $\hat{t}$ is
  - $P(\hat{t} = 30) = \frac{1}{3}$
  - $P(\hat{t} = 40) = \frac{1}{3}$
  - $P(\hat{t} = 50) = \frac{1}{3}$

Mean

- Using the sampling distribution
  - $\mathrm{E}(\hat{t}) = \frac{1}{3} \times 30 + \frac{1}{3} \times 40 + \frac{1}{3} \times 50 = 40.$
- Using the sample probabilities
  - $\mathrm{E}(\hat{t}) = \frac{1}{6} \times 30 + \frac{1}{6} \times 40 + \frac{1}{6} \times 30 + \frac{1}{6} \times 50 + \frac{1}{6} \times 40 + \frac{1}{6} \times 50 = 40.$

Bias

- $t = 5 + 10 + 15 + 10 = 40$
- $\mathrm{E}(\hat{t}) = 40$
- $\mathrm{Bias}(\hat{t}) = \mathrm{E}(\hat{t}) - t = 40 - 40 = 0$
- $\hat{t}$ is *unbiased*.

Example Continued

Sample design 3

- $P(S_1) = P(S_2) = P(S_3) = \frac{1}{3}$
- $Y_1 = 5$; $Y_2 = 10$; $Y_3 = 15$; $Y_4 = 10$
- For $S_1 = \{1, 2\}$, $\hat{t} = 4 \times 7.5 = 30$.
- For $S_2 = \{1, 3\}$, $\hat{t} = 4 \times 10 = 40$.
- For $S_3 = \{1, 4\}$, $\hat{t} = 4 \times 7.5 = 30$.
- $\mathrm{E}(\hat{t}) = \frac{2}{3} \times 30 + \frac{1}{3} \times 40 = 33.33$
- Bias is $33.33 - 40 = -7.67$.

Variance and standard deviation

- $\mathrm{Var}(\hat{t}) = \mathrm{E}(\hat{t} - \mathrm{E}(\hat{t}))^2$
- For sample design 1,

$$\mathrm{Var}(\hat{t}) = \frac{1}{3}(30-40)^2 + \frac{1}{3}(40-40)^2 + \frac{1}{3}(50-40)^2 = 200/3 = 66.67$$

- The standard deviation is $\sqrt{66.67} = 8.2$

Design 3

- For sample design 3,

$$\mathrm{Var}(\hat{t}) = \frac{2}{3}(30-33.33)^2 + \frac{1}{3}(40-33.33)^2 = 22.22$$

- The standard deviation is $\sqrt{22.22}$.
- Design 3 has smaller SD (4.7) than Design 1 (8.2).
- But it is biased.

Example Continued

Mean squared error

- Among unbiased designs, the one with the smallest variance (SD) is best.
- To compare designs (including biased and unbiased designs) in general, we look at the *mean squared error (MSE)*.
- $MSE = E(\hat{t} - t)^2$ ($\hat{t}$ is the random value.)

MSE, variance, and bias

- $MSE = Var + (Bias)^2$
- For design 1, $MSE = Var + 0 = 66.67$
- For design 3, $MSE = 22.22 + (-7.67)^2 = 22.22 + 58.66 = 80.99$

Comparison of designs

- Design 3 has a smaller variance than Design 1.
- Design 3 has a larger MSE than Design 1.
- Therefore, Design 1 is better.
- Designs with small bias can have smaller MSE and be better than unbiased designs.

Simple Random Sample

- A Simple Random Sample (SRS) is a sample where every possible subset of sampling units has the same probability of being the sample.
- Two variants:
  - Simple Random Sampling with Replacement
  - Simple Random Sampling without replacement.

- Sample mean (y_{sample}) is unbiased estimator for population mean

Sample Mean and Variance

Standard error of the mean

- This is the standard deviation of the sampling distribution of the sample mean.
- $\text{Var}(\bar{y}_{sample}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$,
  - where S^2 is population variance $S^2 = \frac{1}{N-1}\sum_{pop}(Y_i - \bar{Y}_{pop})^2$
- The standard error is the square root of the variance.

Finite population correction (fpc)

- The usual formula for the standard error of the mean does not have the term $\left(1 - \frac{n}{N}\right)$.
- This is the finite population correction.
- Note that if n is small, relative to N, the correction is negligible.

Estimation of the standard error of the mean

- Replace the population variance S^2 with the sample variance s^2 in the formula $\text{Var}(\bar{y}_{sample}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$
- And take the square root.

Systematic Sampling

- Proxy for SRS
- Used when no list of population exists
- Let N and n be size of list and sample respectively
  - K be best integer after n/N
  - R is random between 1 and K
  - Sample is R, R+K, R+2K….R + (n-1)K
- Not same as SRS
  - Not possible to have consecutive units in sample
- Bad if population list has some repeating or cyclical patterns
- May be better if list is in increasing or decreasing order.

When should a SRS be used

- Easy to design and infer
- Easier to defend
- No extra information is required
  - If extra information is present, better sampling methods can be designed

Stratified Sampling

- We divide the population into *H* subpopulations called *strata*.
- Each sampling unit is in exactly one stratum.
- We draw independent probability samples from each stratum.
- We pool the information to obtain population estimates.
- Why do we need Stratified Sampling
  - We want to avoid taking a really bad sample.
    - We could take separate SRS of females and males.
  - We want data of known precision for subgroups.
    - Suppose females and males are not be equally represented in the population but we want the same precision for each estimate.
  - A stratified sample may be more convenient to administer and cheaper.
    - Different methods for different strata
  - Stratified sampling can lead to estimates with smaller standard errors compared with an SRS with the same total number of observations.

Stratification

- If there is a choice, which stratification variable should be used? By stratification variable is meant the characteristics used for subdividing the population into strata.
  - For example, would an age and sex stratification be preferred to a stratification by occupational groups?
  - Feature Selection
- How should strata be demarcated? If the stratification uses age groups, what age intervals should be used to set up the strata?
  - *Clustering*
- How many strata should there be? How many age groups should there be, if age is a stratification variable?
  - Binning or Discretization
- A sampling design and a sample size must be specified in each stratum. Often the same type of sampling design is applied in all of the strata.
- An estimator must be specified for each stratum. Often this choice is also made uniformly for all strata.

Notation

- Strata are labeled 1 to H.
- N_h is population size for stratum h.
- $N = N_1 + N_2 + \ldots + N_H$
- Stratified random sampling is the simplest form of stratified sampling; we take of size n_h from each stratum.

Population quantities

- $y_{h,j}$ is the value of the jth unit in stratum h.
- t_h is the total for stratum h.
- t is the population total.
- $\bar{y}_{h,U}$ is the population mean in stratum h.
- $\bar{y}_U$ is the overall population mean.
- S_h^2 is the population variance in stratum h (uses $N_h - 1$ in the denominator).

Sample quantities

- $\bar{y}_h$ is the sample mean for stratum h.
- $\hat{t}_h$ is the sample estimate of the total for stratum h.
- s_h^2 is the sample variance in stratum h.

Estimates of population parameters

$$\begin{aligned} \hat{t}_{str} &= \sum_{strata} \hat{t}_h \\ \bar{y}_{str} &= \hat{t}_{str}/N \end{aligned}$$

Variance

- The variance of $\hat{t}_{str}$ is the sum of the variances of the $\hat{t}_h$.
- The variance of $\hat{t}_h$ is obtained using the methods for SRS's.

Quota Sampling

- Stratification performed
- Individual Strata not sampled as per probability sampling
  - Solely on discretion of interviewer
  - Used when probability sampling is expensive.
- Drawback
  - Selection Bias
  - Estimates may be unbiased
  - No way of even knowing that

Cluster Sampling

- A cluster is a group of observation units (or “elements”)

Population	Obs Unit	Cluster
U.S. residents	person	household
Tucson households	household	city block, or postal route
UA employees	employee	department
Maple trees in Vermont	tree	1 km × 1 km plot

- A cluster sample is a probability sample in which a sampling unit is a cluster

Frame	SU	OU
List of phone numbers	phone number	person
List of blocks	block	household
List of UA departments	department	faculty member
List of plots	plot	tree

Cluster Sampling

Why use cluster samples?

- Constructing a frame of the observation units may be difficult, expensive, or impossible.
  - Customers of a store
  - Birds in a region
- The population may be widely distributed geographically or may occur in natural clusters.
  - Residents of nursing homes
  - College students in dorms/classes

Comparison with stratification

- With both clusters and stratification we partition the population into subgroups (strata or clusters).
- With stratification, we sample from *each* of the subgroups.
- With cluster sampling, we sample *all* of the units in a *subset* of subgroups.

Cluster Sampling

- Typically larger variance
- Estimates simple in case of equal size clusters
- Cluster of unequal size
  - No new ideas
  - Formulas are messy

- This is one-stage cluster sampling
- Two stage cluster sampling
  - First select clusters
  - Sample within cluster
  - Two sources of error
    - Cluster selection
    - Sample selection

Survey Sampling Issues

- Non Response and Missing Value
  - Entire Unit is Missing
  - Some questions are not answered
- Few options
  - Should try to prevent it by design
  - Use model to predict and fill missing values
  - Ignore missing values – Most Common
- Implicitly assumes that the non respondents are similar to the respondents; similarly for items
- Evidence suggests that this is not true in general.
- Results reported should be taken as estimating the population of respondents.
- Report the non response rate.

Non Response

- Survey content
  - personal info
  - helped by reordering
- Time of survey
  - Call during Break?
- Interviewers
- Data-collection method
  - telephone vs mail vs personal
- Questionnaire design
  - appeal to cognitive science
- Respondent burden
- Survey introduction
  - why data used
  - assured confidentiality
- Incentives and *disincentives*
- Follow-up
- The attitude of management toward non response.

Second Part

W a r m -up: Distinct Elements

Theorem: If a table contains at least k distinct elements of frequency at least ε, then, for any fixed $0 < \delta < 1$, a sample of size $\frac{1}{\varepsilon}\log\frac{k}{\delta}$ detects k such elements with a prob. at least $(1 - \delta)$.

Proof : Let $a_1, \ldots, a_k$ be such elements. Let Sample size be S. If we guarantee that each element a_i is missed with a probability of utmost $\frac{\delta}{k}$, then, we get the desired guarantee.

$\therefore$ We want to ensure that : $(1-\varepsilon)^S \leq \frac{\delta}{k}$.

Let $S = \frac{1}{\varepsilon}\log\frac{k}{\delta}$

$$\left((1-\varepsilon)^{\frac{1}{\varepsilon}}\right)^{\log\frac{k}{\delta}} \leq \left(\frac{1}{e}\right)^{\log\frac{k}{\delta}} \leq \frac{\delta}{k}$$

Callaghan, Mishra, Meyerson, Guha, Motwani (ICDE 02)

Case Study: Sampling Techniques for Association Rule Mining

Frequent Itemset Mining (FIM)

- A set of m items $I = \{I_1, I_2, \ldots, I_m\}$
- N transactions $t_1, t_2, \ldots t_n$
- Each transaction t_i is a subset of I
- An *itemset* X is a subset of I

$$I = \{1,2,3,4,5,6\}$$
$$t_1 = \{2,5\}$$
$$t_2 = \{1,4,6\}$$
$$t_3 = \{1,5,6\}$$

$$frequency(X) = \frac{\#\,\text{transactions containing X}}{\text{Total number of transactions (N)}}$$

Problem: Given a parameter $0 \le \theta \le 1$, find all itemsets that are θ-frequent.

An itemset is θ-frequent if its frequency is $\ge \theta$

Since we are studying sampling, we will use *1/h* to denote the upperbound on the desired probability of failure. Here *h* is a suitably large integer.

Association Rule Mining (ARM)

- A set of m items I = $\{I_1, I_2, \ldots, I_m\}$
- N transactions $t_1, t_2, \ldots t_n$
- Each transaction t_i is a subset of I
- An *itemset* X is a subset of I

$$I = \{1,2,3,4,5,6\}$$
$$t_1 = \{2,5\}$$
$$t_2 = \{1,4,6\}$$
$$t_3 = \{1,5,6\}$$

$$frequency(X) = \frac{\#\,\text{transactions containing X}}{\text{Total number of transactions (N)}}$$

Problem: Given parameters $0 \le \theta \le 1$ (called *threshold*), and $0 \le \gamma \le 1$ (called *confidence*), find association rules of the following type:

A → B Where B is a superset of A and the rule implies that,

((frequency(A) ≥ frequency(B) ≥ θ) AND (frequency(B)/frequency(A) ≥ γ))

Algorithms for Association Rule Mining

- Level-wise discovery of frequent itemsets (Apriori – AS94 is the main reference)
  - Observation: All itemsets of a frequent itemsets have to be frequent.
  - Level-by-level construction starting with the frequent singletons.
- Partition-based discovery of frequent itemsets (SON95)
  - Partition the database into small enough chunks
  - Compute frequent items on each chunk and their union is the candidate list.
  - Candidate list is pruned by a second pass of the database.
- Highly inefficient requiring a minimum of two passes of the entire database.

- In emerging applications, even a single pass is undesirable
  - Number of yearly transactions across Walmart branches is to the tune of 10^{10}.
  - The number of records in scientific experiments (CERN databases) orders of magnitudes larger than the commercial applications!
- So, working with smaller samples is absolutely critical for carrying out association rule mining.

Sampling Techniques for Association Rule Mining

- Intuitively it is clear that results of ARM on small samples should closely match the results on the entire database.
- Several papers treating it as a heuristic for efficiency...
  - Perhaps, the work of T96 (Toivonen in VLDB 96) is the first that studied the effectiveness of sampling for frequent itemset mining.
  - Closely following this work was the work of ZPLO97 (Zaki, Parthasarathy, Li, and Ogihara in KDD 97) that developed rules-of-thumb for sampling.
- Two recent works have given further understanding of sampling without any pass of the database
  - PRUV10 (Pietracaprina, Riondato, Upfal, Vandin in ECML PKDD 2010) presented a framework to study the related problem of discovering top-K frequent itemsets.
  - CPS09 (Chakaravarthy, Pandit, Sabharwal in ICDT 2009) presented a comprehensive theoretical framework for studying sampling in the context of ARM and presented surprising bounds that are independent of database size!

Chernoff Bounds

- Let X be a random variable obtained by summing up a set of *n* independent, identical, indicator variables. The following Chernoff bounds on the deviation of X from its expected value are well known.

$$\Pr[X \leq (1-\delta)\mathrm{E}[X]] \leq e^{-\left(\frac{\delta^2 \mathrm{E}[X]}{2}\right)} \quad \text{(CH1)}$$

$$\Pr[X \geq (1+\delta)\mathrm{E}[X]] \leq e^{-\left(\frac{\delta^2 \mathrm{E}[X]}{3}\right)} \quad \text{(CH2)}$$

$$\Pr[|X - \mathrm{E}[X]| \geq \delta] \leq 2e^{-\left(\frac{2\delta^2}{n}\right)} \quad \text{(CH3)}$$

Example of applying Chernoff bounds for Frequencies

Let S be a random sample drawn from a database. Let x be an itemset of interest. Let f_x be its frequency in the database. Let f_x^S be its frequency in the sample. Let us bound $\Pr[|f_x - f_x^S| \geq \varepsilon]$.

$$\Pr[|X - E[X]| \geq \delta] \leq 2e^{-\frac{2\delta^2}{n}}$$

Here, $n = |S|$. $E[X] = f_x|S|$. $X = f_X^S|S|$.

$$\Pr[|f_x^S - f_x| \geq \varepsilon] = \Pr[|f_x^S|S| - f_x|S|| \geq \varepsilon|S|] \leq 2e^{-\frac{2\varepsilon^2|S|^2}{S}}$$

$$\therefore \Pr[|f_x^S - f_x| \geq \varepsilon] \leq 2e^{-2\varepsilon^2|S|}$$

First Cut Idea – T96

- Use a small sample to get the list of possible frequent itemsets
- Follow it up with a single scan of the entire database to filter only true frequent itemsets

- Problem: How do you ensure that all possible frequent itemsets are identified from the sample?
- Idea: Run Apriori with a lower threshold than the specified threshold.

- If the sample size is *s*, and threshold is θ, then, setting lower threshold θ' and we want to identify a frequent itemset with a probability of atleast (1 – 1/h), it is sufficient to keep

$$\theta' \leq \left(1 - \sqrt{\frac{2\log h}{\theta s}}\right)\theta$$

Getting used to these calculations

- In a sample of size s, a θ - frequent itemset is expected appear θs times. That is E[X].
- We want to find an ε such that, $\Pr[X < (1-\varepsilon)\theta s] < \frac{1}{h}$.

$$\Pr[X < (1-\varepsilon)\theta s] \leq e^{\frac{-\varepsilon^2 \theta s}{2}} \leq \frac{1}{h}$$

$$e^{\frac{\varepsilon^2 \theta s}{2}} \geq h \Rightarrow \varepsilon \geq \sqrt{\frac{2\log h}{\theta s}}$$

$$\therefore \theta' \leq \left(1 - \sqrt{\frac{2\log h}{\theta s}}\right)\theta$$

Algorithm

- Obtain a sample of size s
- Run Apriori with a lower threshold given by the below formula
- Filter the identified itemsets by a single scan of the database
- Experimentally validated the technique on moderate sized databases.
- Objection: Still Requires a single scan of the database ☹

$$\theta' \leq \left(1 - \sqrt{\frac{2\log h}{\theta s}}\right)\theta$$

Heuristic Approach: ZPLO97

- The Chernoff Bound for ensuring that a given frequent itemset appears in the sample itself was deemed infeasible!
  - Mainly because the experiments till then worked with smaller databases
- So, instead developed rule-of-thumb approach to discover rules
  - Main Goal: To avoid scan of the database

- Empirical Study of Sampling
  - Toy Databases by today's standards and some synthetic datasets.
  - Mainly considered the option of sampling a fixed % of the database: varying from 1% to 25%
  - Studied accuracy, and gains in efficiency in great detail
  - Concluded that sampling range of 10%-20% of the database size is required to obtain acceptable accuracy and gains in efficiency

- Except for occasional works that tried adaptively detecting when to stop sampling, no major insights were obtained.

An Ideal Sampling Based Solution

Would like to analyze only a small sample of the database and be able to

- Report all θ -frequent itemsets.
- Do not report any itemset of frequency $< \theta$
- No subsequent pass over the database.

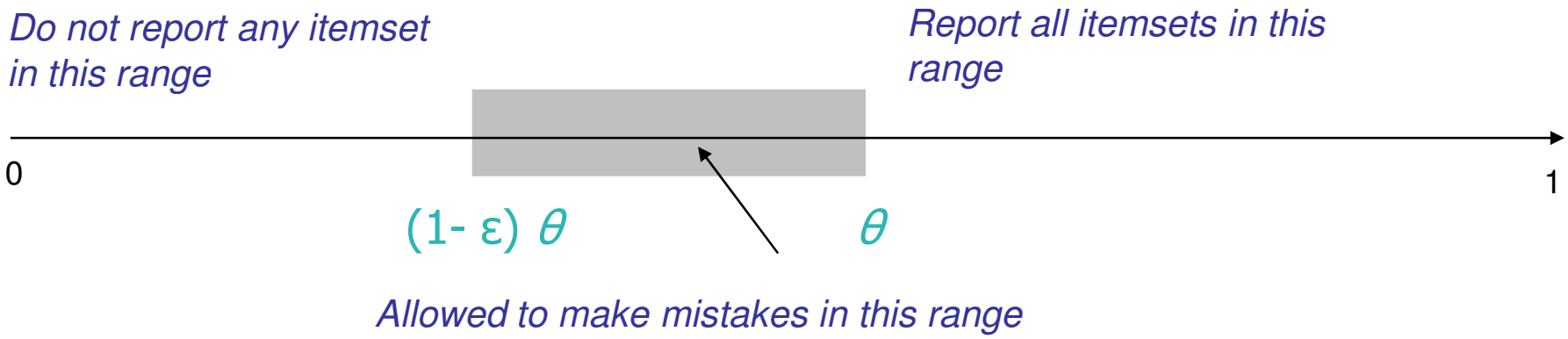
What should be the sample size?

- Is this even possible?
- Boundary Case: an itemset with frequency really close to θ
- Such an itemset cannot be classified based on the sample.

ε-close Frequent Itemset Mining: CPS09

Problem: Given parameters $0 \le \theta \le 1$, $0 \le \varepsilon \le 1$, report all itemsets that have frequency of at least θ and do not report any itemset whose frequency is below $(1-\varepsilon)\,\theta$.

Note: The algorithm may end up reporting some itemsets whose frequency is in the range of $((1-\varepsilon)\,\theta,\ \theta]$



Main Question : What is the sample size required ?

Main Results

Naive Bound :

$$S \geq \frac{12}{\varepsilon^2 \theta} \min \begin{Bmatrix} m + \log h, \\ \log N + \log h + \Delta \end{Bmatrix}$$

Better Bound :

$$S \geq \frac{24}{\varepsilon^2 (1-\varepsilon)\theta} \left[\Delta + 5 + \log \frac{5h}{(1-\varepsilon)\theta} \right]$$

m – number of items

h - probability of error

Δ - Bound on the number of items in a transaction

N – Number of transactions

ε - closeness parameter

θ - Required support

- Note: The bound is Independent of m and N!

- Let the improved bound be denoted by B.

Why is it important?

- Theoretical understanding of the sampling technique for the Association Rule Mining Problem.
- The result starts kicking-in when the database sizes grow.
  - We already saw examples of such massive databases
- Our result implies that a sample of just 5 million is sufficient to solve the problem with desired level of accuracies on databases with 10 Billion transactions.
- A follow-up empirical work with Prof. Jayant Haritsa of IISc has found that much smaller samples are sufficient! → Can we prove better bounds?!

Sampling Algorithm

- Sampling Algorithm
  - Pick a random sample of size S
  - Report any itemset with frequency $(1-\varepsilon/2)\theta$ in the sample

- A sample is said to be *good*, if
  - reports every itemset with frequency $\geq \theta$
  - does not report any itemset with frequency $\leq (1-\varepsilon)\theta$
- *Bad,* otherwise

Goal : Choose sample size S such that

$$\Pr[\text{ Sample is bad }] \leq 1 / h$$

Analysis

Applying Chernoff Bounds

- W be an itemset with frequency ≥ θ

$$\Pr[\text{W is not reported}] \leq e^{-\frac{\varepsilon^2 S\theta}{8}}$$

- W be an itemset with frequency ≤ (1-ε) θ

$$\Pr[\text{W is reported}] \leq e^{-\frac{\varepsilon^2 S\theta}{12}}$$

$$\text{If } S \geq \frac{12}{\varepsilon^2\theta}\log h, \text{ then, above Pr} \leq \frac{1}{h}$$

The Naïve Bound

Number of occurant itemsets $\leq 2^m$

Number of occurant itemsets $\leq N2^{\Delta}$

$$\Pr[\text{ Sample is bad }] \leq \min\left\{2^m, N2^{\Delta}\right\} e^{-\frac{\varepsilon^2 S\theta}{12}}$$

$$S \geq \frac{12}{\varepsilon^2\theta}\left[\min(m, \log N + \Delta) + \log h\right]$$

Number of θ-frequent itemsets : A closer look

- How many θ-frequent itemsets are possible?

Naive answer : $\min\{2^m, N2^{\Delta}\}$

A better bound : $\frac{2^{\Delta}}{\theta}$

- This observation is sufficient to show that all the itemsets in the RHS will be reported if the sample size is B.

- We have to do more work to show that the itemsets from LHS will not be reported.

Accepting Frequent Itemsets

- How many θ-frequent itemsets are possible? $\frac{2^{\Delta}}{\theta}$

$$\Pr[\text{ Some } \theta \text{ - frequent itemset is not reported }] \leq \frac{2^{\Delta}}{\theta} e^{-\frac{\varepsilon^2 S\theta}{12}}$$

$$\text{If } S \geq \frac{12}{\varepsilon^2\theta}\left[\Delta + \log h + \log\frac{1}{\theta}\right],$$

$$\Pr[\text{ Some } \theta \text{ - frequent itemset is not reported }] \leq \frac{1}{h}$$

Rejecting Non-frequent Itemsets

For any $0 \leq \beta \leq 1$,

$$\text{Number of } \beta\text{ - frequent itemsets} \leq \frac{2^{\Delta}}{\beta}$$

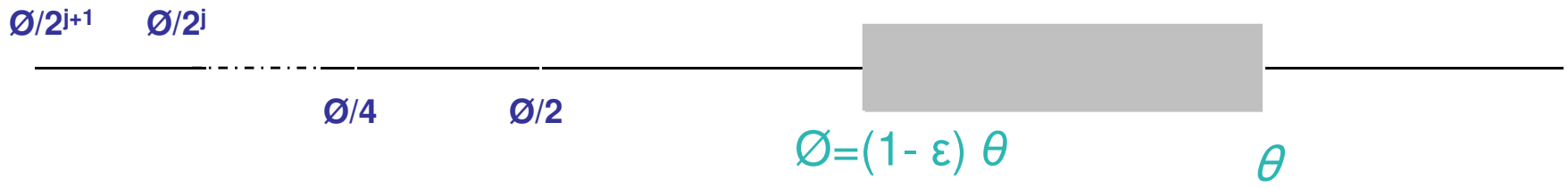
- This bound is useful for a constant β.
  - Not useful for itemsets appearing in O(N) transactions

Main insight of the analysis

- Consider a particular frequency f in the range (1/N, (1- ε) θ); as f approaches 1/N,
  - Number of f-frequent itemsets increases
  - Probability that an itemset with frequency <= f is reported by the sampling algorithm decreases.
- Insight: the rate at which the number of itemsets increases is slower than the rate at which the probability decreases.

Rejecting Non-frequent Itemsets

Geometrically divide the range



Fix a range $R_j = [Ø/2^{j+1}, Ø/2^j]$

$$\text{Number of itemsets whose frequency lies in } R_j \leq \frac{2^{\Delta + j + 1}}{\phi}$$

For any itemset W whose frequency lies in the range R_j

$$\Pr[\text{ W is reported }] \leq 2^{-(j-2)\,\phi\, S/2}$$

$$\text{Taking } S \geq \frac{2}{\phi}\left[\Delta + \log\frac{h}{\phi} + 5\right]: \quad \Pr[\text{Failure in } R_j] \leq \frac{2^{-(j-2)}}{h}$$

Rejecting Non-frequent Itemsets

Taking $S \geq \frac{2}{\phi}\left[\Delta + \log\frac{h}{\phi} + 5\right]$: $\Pr[\text{Failure in } R_j] \leq \frac{2^{-(j-2)}}{h}$

Summing over all ranges

$$\Pr[\text{Failure}] \leq \sum_j \frac{2^{-(j-2)}}{h} \leq \frac{1}{h}$$

Theorem :

With a sample of size $S \geq \frac{24}{\varepsilon^2(1-\varepsilon)\theta}\left[\Delta + 5 + \log\frac{5h}{(1-\varepsilon)\theta}\right]$,

we can solve the ε-close frequent itemset mining,

with a probabilit y $\geq 1 - \frac{1}{h}$

Association Rule Mining

- Association rule: A rule of the form W ➔I, where
  - W is an itemset and I is an item
  - Intuition: whenever W is present in a transaction, I is likely to be present as well.

Problem: Given a *support* θ and a *confidence* γ, find association rules of the type W➔I such that

- $W \cup I$ is θ - frequent
- $\dfrac{frequncy\ \ (W \cup I)}{frequency\ \ (W)} \geq \gamma$

- ε-close association rule mining,
  - output all rules having support θ and confidence γ
  - do not output any rules having support less than $(1-\varepsilon)\theta$
  - do not output any rules having confidence less than $(1-\varepsilon)\gamma$

Sampling Algorithm

- Sampling Algorithm
  - Pick a random sample of size S
  - Find F, the set of all itemsets with frequency $(1-\varepsilon/2)\theta$ in the sample
  - Output all association rules of the form W→I such that
    - W U I is in F
    - The rule has a confidence of at least $\left(\frac{1-\varepsilon/4}{1+\varepsilon/4}\right)\gamma$

Sample Size Bound

- Support is already dealt with; have to take care of only confidence
  - Have to ensure that both numerator and denominator values in the sample do not deviate much from their original value
  - We apply Chernoff bounds to ensure this.
    - Since both numerator and denominator are θ-frequent, the number of such sets is bounded
    - therefore sample size required for Chernoff bounds is small

Theorem :

With a sample of size $S \geq \frac{48}{\varepsilon^2(1-\varepsilon)\theta}\left[\Delta + 5 + \log\frac{5h}{(1-\varepsilon)\theta}\right]$,

we can solve the ε - close association rule mining,

with a probability $\geq 1 - \frac{1}{h}$

Experimental Results

- Initial experimental results on datasets obtained from the standard synthetic data generator (IBM Quest).
- Experimentation with datasets of size 100 Million and 250 Million transactions.
- Average Transaction Length = 10
- Maximum Transaction Length, $\Delta = 35$.
- $\varepsilon = 0.1$
- $h = 0.9$
- Threshold $\theta = 0.02$ and $\theta = 0.01$
- Suggested Sample Bounds = 6 Million and 12 Million respectively.

Experimental Results: Threshold, θ : 0.02

Sample Size	% of Db	N_{FN}	N_{AFP}	N_{UFP}
6 Million (Bound)	5	0	126	0
3 Million	2.5	0	161	0
0.5 Million	1.25	0	167	0
0.1 Million	0.04	0	135	0
0.04 Million	0.1	0	159	1
0.025 Million	0.04	8	152	14

True Freq. Items
1268

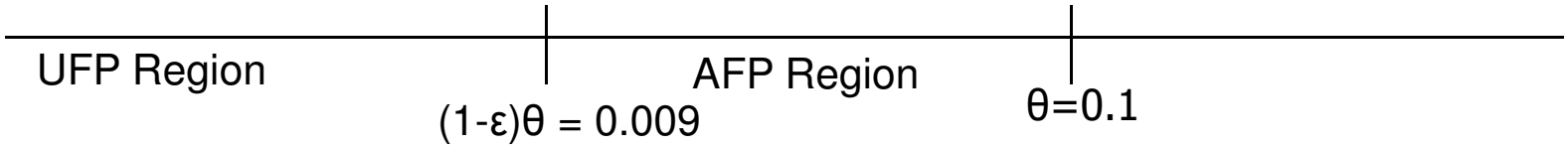
N_{FN} = # (False Negatives); N_{AFP} = #(Acceptable False Positives) ;N_{UFP} = #(Unacceptable False Positives)

Experimental Results: Threshold, θ : 0.01

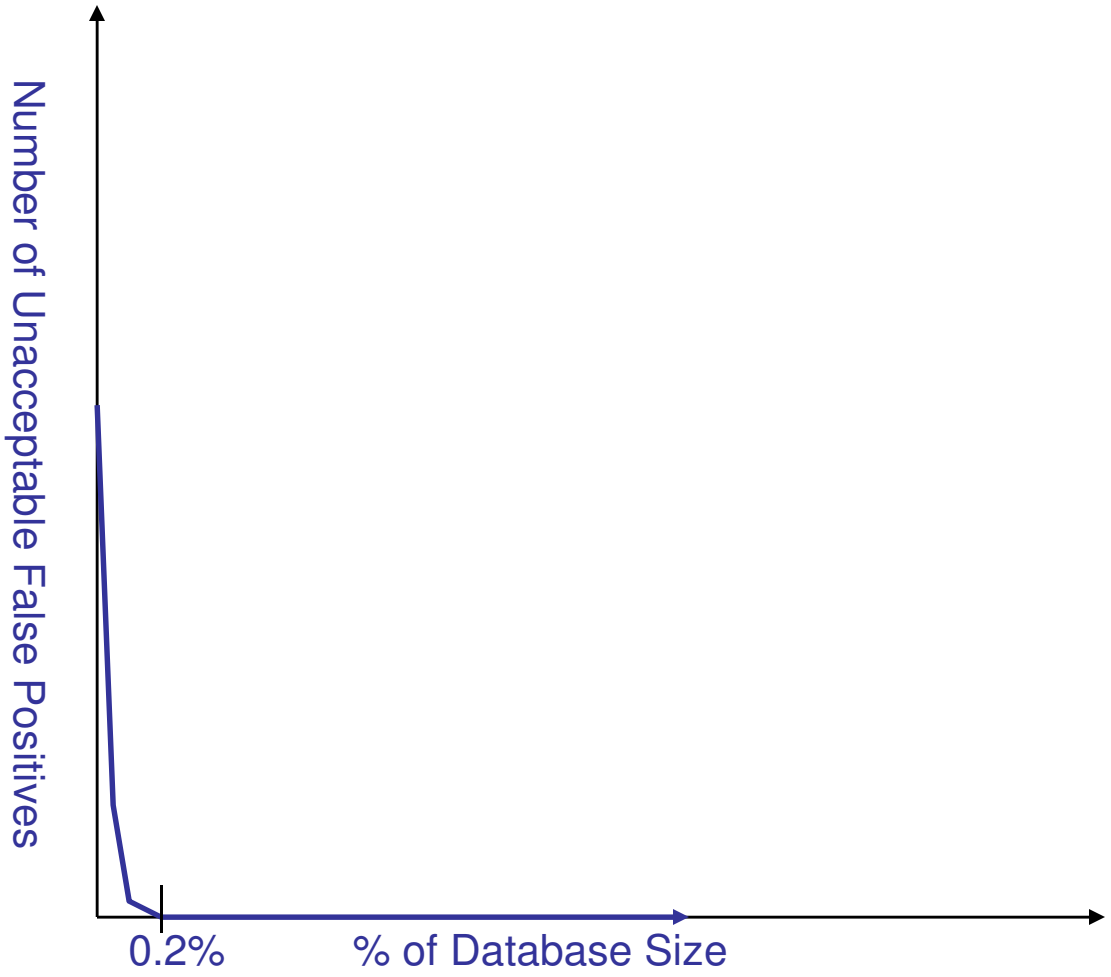
Sample Size	% of Db	N_{FN}	N_{AFP}	N_{UFP}
12 Million (Bound)	5	0	460	0
6 Million	2.5	0	474	0
3 Million	1.25	0	486	0
0.5 Million	0.2	0	492	0
0.3 Million	0.1	0	462	1
0.1 Million	0.04	5	573	19

True Freq. Items 4837

N_{FN} = # (False Negatives); N_{AFP} = #(Acceptable False Positives) ;N_{UFP} = #(Unacceptable False Positives)



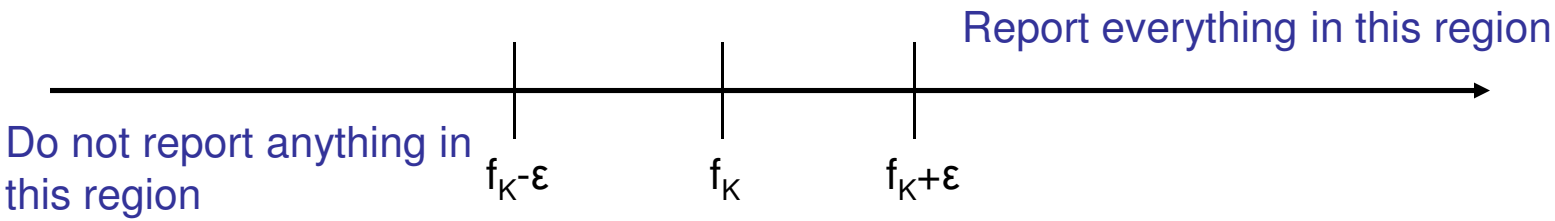
Number of Unacceptable False Positives



Mining top-K frequent itemsets

Pietracaprina, Riondato, Upfal, Vandin in ECML PKDD 2010 (PRUV10)

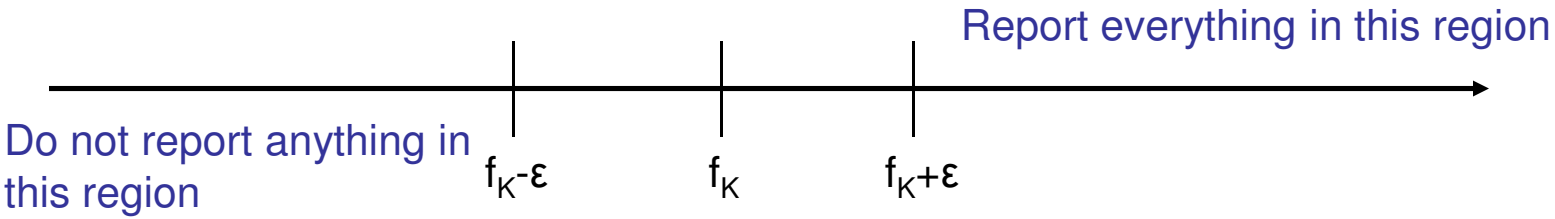
- Considered the problem of sampling for obtaining top-K frequent itemsets by frequency.
- They consider the problem of returning only those itemsets whose size is bounded by a constant "w". Let f_K be the frequency of the Kth such frequent itemset.
- You need to ensure that the frequency of an itemset in the sample does not deviate from its original frequency by more than ε.



Mining top-K frequent itemsets

- Considered the problem of sampling for obtaining top-K frequent itemsets by frequency.
- They consider the problem of returning only those itemsets whose size is bounded by a constant "w". Let f_K be the frequency of the Kth such frequent itemset.
- You need to ensure that the frequency of an itemset in the sample does not deviate from its original frequency by more than ε.

- CPS09 gives us the following bound:

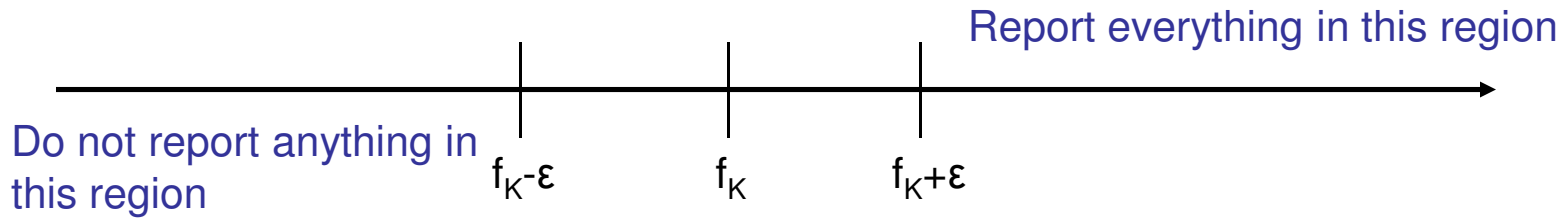


$$S \geq \frac{24 f_K^2}{\varepsilon^2(\theta - \varepsilon)}(\Delta + 5 + \log 5h)$$

Problem: We do not know f_K!

Mining top-K frequent itemsets

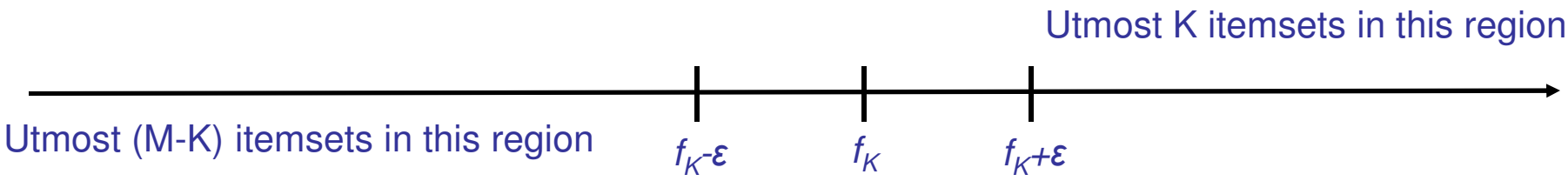
- Considered the problem of sampling for obtaining top-K frequent itemsets by frequency.
- They consider the problem of returning only those itemsets whose size is bounded by a constant "w". Let f_K be the frequency of the Kth such frequent itemset.
- You need to ensure that the frequency of an itemset in the sample does not deviate from its original frequency by more than ε.



- They show a bound which only depends on K and "M". "M" depends on "m", the number of items, as follows. Essentially, M is the set of all itemsets of size utmost w.

$$M = \sum_{i=0}^{w} \binom{m}{i}$$

Main Idea



Therefore, there are K(M-K) pairs that need to get separated.
There are utmost M itemsets for which absolute deviation needs to be respected.

Therefore there are utmost (2M+K(M-K)) events for which we need to ensure correctness. If each one of them satisfies the correctness with probability given below, we can happily apply union bound and be done.

$$\frac{1}{h(M + K(M - K)}$$

Calculations

Suppose S is the size of the sample, then, it needs to be such that:

$$e^{-\left(\frac{\varepsilon^2 S}{2}\right)} \leq \frac{1}{h(2M + K(M-K))}$$

$$e^{\left(\frac{\varepsilon^2 S}{2}\right)} \geq h(2M + K(M-K))$$

$$\frac{\varepsilon^2 S}{2} \geq \log(h(2M + K(M-K)))$$

$$S \geq \frac{2}{\varepsilon^2}\log(h(2M + K(M-K)))$$

Mining top-K frequent itemsets

Theorem: For fixed, 0≤ε≤1, and a constant *h*, a sample of B random transactions solves the approximate top-K frequent itemsets with a probability of at least (1-1/*h*) if B satisfies the following:

$$B \geq \frac{2}{\varepsilon^2} \ln(h(2M + K(M - K))$$

Mining top-K frequent itemsets: Experimentation

- They experimented with 3 datasets.
- *T10I4D100K* – Artificially created dataset with IBM Quest tool with 100K transactions, 1000 items, and average transaction length of 19.
- *kosarak*, a click-stream dataset of a Hungarian website with 41,270 items, 1 Million transactions, and average transaction length of 8.
- *webdocs*, a special dataset that views web documents as “transactions”. Each document is a transaction containing its words. There are 1.7 Million transactions (documents) and there are 0.52 Million items (words), and average transaction length of 177.
- ε=0.02 and h = 10.
- Experimental Results:
  - In 85% of the runs, all the true top-K frequent itemsets were reported and no itemset in the false positive region were reported.
  - More importantly, in 15% of the experiments, upto 5% of true top-K itemsets were missed (in accordance with the problem definition).

Frequent Items Mining on the Streaming Model

- Let θ be the frequency threshold as before and ε, the tolerance parameter as before.
- But, the data is coming in a *stream*
  - Huge number of items are coming one-by-one.
  - There is not enough memory to keep all the seen items.
  - We cannot even recall an item that was seen in the past and not kept in the sample.
- As before, we want to identify θ-frequent items.
- Main Question: What is the size of the sample required to identify the θ-frequent items?
- As before, it turns out that the sample size required does not even depend on the number of items in the stream.
  - We will present a nice algorithm of Manku and Motwani, MM02 (in VLDB 02)

Frequent Items Mining: Data Structure

- Let *S* be the size of the sample that the algorithm keeps at every point in time. The sample is stored in the following data structure.

Element	Count
e	$count_e$

Frequent Items Mining: Algorithm.

- Let *S* denote the sample size at any point in time.
- The algorithm runs in phases and parameterized by t.

- First phase is special. It looks at first 2t elements and initializes the table.

Element	Count
e	$count_e$

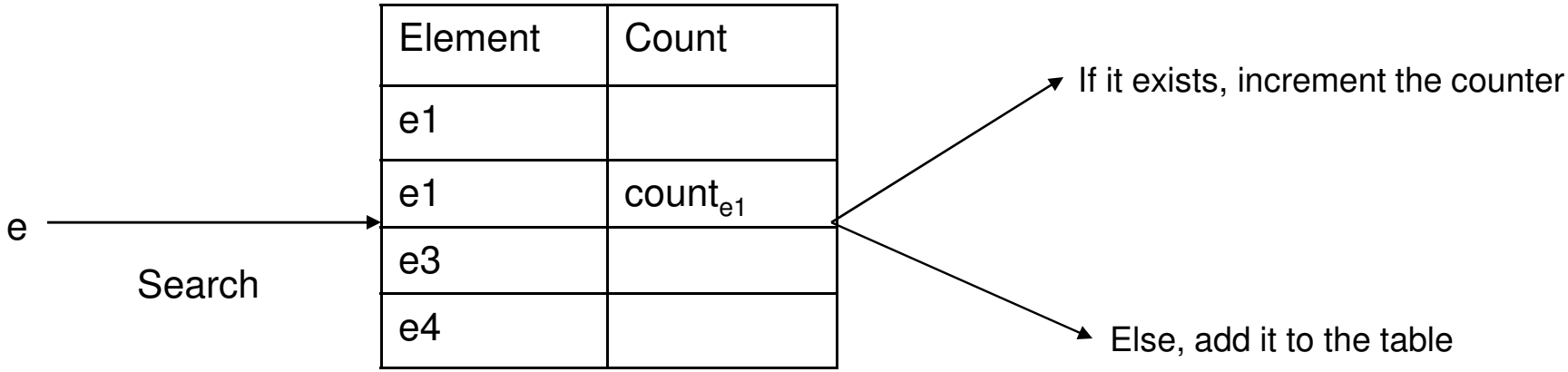
Frequent Items Mining: Algorithm.

- Let *S* denote the sample size at any point in time.
- The algorithm runs in phases and parameterized by t.

- First phase is special. It looks at first 2t elements and initializes the table.

- Phase i begins after $(2^{(i-1)} -1)t$ items have appeared on the stream.
- At phase i, selectivity is $r = 2^{(i-1)}$.

- In phase i, when an item *e* appears on the stream, we do the following:
  - If e exists in the sample, we increment its counter by one.
  - If it does not exist, then, we add it to the sample with a probability of *1/r* and initialize its counter to 1.



Frequent Items Mining: Algorithm.

- Let *S* denote the sample size at any point in time.
- The algorithm runs in phases and parameterized by t.

- First phase is special. It looks at first 2t elements and initializes the table.

- Phase i begins after $(2^{(i-1)} - 1)t$ items have appeared on the stream and it sees $2^{(i-1)}$ items
- At phase i, selectivity is $r = 2^{(i-1)}$.

- In phase i, when an item *e* appears on the stream, we do the following:
  - If e exists in the sample, we increment its counter by one.
  - If it does not exist, then, we add it to the sample with a probability of *1/r* and initialize its counter to 1.

- At the end of phase i, when we are setting r to 2^i, we do the following for each entry $(e, count_e)$ in the sample *S*
  - Toss a fair coin till we get a HEADS.
  - Before that, every time we get TAILS, we decrement by $count_e$ one. If the count reaches zero, we delete the entry for *e*, and move on to the next element.

Frequent Items Mining: Algorithm.

- Last step

Element	Count
e	$count_e$

- If we get $count_e$ TAILS, we delete e
- Else, if we get *x* TAILS and then a HEADS, we set $count_e = count_e - x$;

Theorem

Theorem : When we run the algorithm with $t = \frac{1}{\varepsilon}\log\left(\frac{1}{\theta} \bullet \frac{1}{\delta}\right)$, we get an ε - close solution with a probability of at least $(1 - \delta)$ with the expected number of entries in the sample being 2t.

Expected #(Entries) in the Sample

- During the ith phase
  - Look at $2^{(i-1)}t$ items and each item can result in a new entry with a probability of utmost $1/2^{(i-1)}$.
  - Therefore, new entries are utmost t.
  - So, if we maintain an invariant that after the last step of ith phase, expected number of entries is t, then, at each time, our expected number of entries during all steps is 2t.
- Look at the last step

Element	Count
e	$count_e$

- If we get $count_e$ TAILS, we delete e
- Else, if we get *x* TAILS and then a head, we set $count_e = count_e - x$;

- This is equivalent to sampling all the seen items with a probability equal to new r, $r = 1/2^i$
- Therefore, expected entries $\geq (2^i t)(1/2^i) = t$.

- So, the invariant we are looking to satisfy is indeed satisfied and the number of entries is bounded by 2t in expectation.

Probability of Success

- Let N be the number of items in the stream. It is easy to see that:
  - $rt \leq N \Rightarrow 1/r \geq t/N$.
- Also easy to see that, number of frequent items is utmost $1/\theta$. So, probability of failure for each element should be utmost $\delta\theta$.

- Consider an element e which is θ-frequent. It will not be reported as θ-frequent if more than εN occurrences of it are not included in the sample.

$$\left(1-\frac{1}{r}\right)^{\varepsilon N} \leq \delta\theta$$

$$\left(1-\frac{t}{N}\right)^{\varepsilon N} \leq e^{-\varepsilon t} \leq \delta\theta$$

$$\varepsilon t \geq \log\left(\frac{1}{\delta\theta}\right)$$

$$t \geq \frac{1}{\varepsilon}\log\left(\frac{1}{\delta\theta}\right)$$